

Parallel, linear-scaling building-block and embedding method based on localized orbitals and orbital-specific basis sets.

Luis Seijo* and Zoila Barandiarán†

Departamento de Química, C-XIV,

Universidad Autónoma de Madrid, 28049 Madrid, Spain

and

Instituto Universitario de Ciencia de Materiales Nicolás Cabrera,

Universidad Autónoma de Madrid, 28049 Madrid, Spain

We present a new linear scaling method for the energy minimization step of semiempirical and first-principles Hartree-Fock and Kohn-Sham calculations. It is based on the self-consistent calculation of the optimum localized orbitals of any localization method of choice and on the use of orbital-specific basis sets. The full set of localized orbitals of a large molecule is seen as an orbital mosaic where each *tessera* is made of only a few of them. The orbital *tesserae* are computed out of a set of embedded cluster pseudoeigenvalue coupled equations which are solved in a building-block self-consistent fashion. In each iteration, the embedded cluster equations are solved independently of each other and, as a result, the method is parallel at a high level of the calculation. In addition to full system calculations, the method enables to perform simpler, much less demanding embedded cluster calculations, where only a fraction of the localized molecular orbitals are variational while the rest are frozen, taking advantage of the transferability of the localized orbitals of a given localization method between similar molecules. Monitoring single point energy calculations of large poly(ethylene oxide) molecules and three dimensional carbon monoxide clusters using an extended Hückel Hamiltonian are presented.

I. INTRODUCTION

The development of linear scaling computational methods for electronic structure calculations in molecules and solids with a very large number of atoms, (i.e. methods whose computational demands grow as the first power of the size of the system,) has been a very active and successful field in the last decade.^{1,2} Linear scaling and low order scaling techniques exist for the computation of the one-electron effective Hamiltonian matrix in first-principles calculations (with density functional theory and wave function based methods),^{3,4,5,6,7,8} as well as for the energy minimization step,^{1,2,9,10,11,12,13,14,15,16,17} which is a common step to first-principles and semiempirical calculations. In low order scaling energy minimization methods, the traditional cubic scaling diagonalization of the matrix representation of the effective one-electron Hamiltonian in a finite basis set,¹⁸ which leads to the canonical orbitals, is substituted by different algorithms which, either solve directly for the unique optimal density matrix,^{9,10,11,16,17} or for some sort of arbitrary optimal localized orbitals.^{5,12,14,15} In parallel to linear scaling electronic structure methods, a significant development has also been made in embedding methods, which focus the computational effort on local properties of a system^{19,20} (see Refs. 21 and 22 for recent reviews).

In this paper, we present a new method for the energy minimization step which can be used in semiempirical and first-principles Hartree-Fock and Kohn-Sham

calculations. It is expected to be useful as a complement of linear-scaling methods for the computation of the Hamiltonian matrix, which are nowadays available for local exchange-correlation potentials and Coulomb potentials,^{3,4,5,6,7} as well as for exact exchange fields.⁸ The energy minimization method is based on exploiting the self-consistent calculation of the optimum localized orbitals of any localization method of choice and the use of orbital-specific basis sets. The n occupied localized orbitals of a very large molecule which correspond to a chosen localization method (e.g. the popular methods of Boys,²³ Edmiston-Ruedenberg,²⁴ and Pipek-Mezey,²⁵ or any other available or newly developed localization method) can be regarded as a mosaic of orbitals, where each of its component *tesserae* contains only a few orbitals. In the present method we define an orbital *tessera* as a subset of the occupied localized orbitals (of the method of choice) which are localized in some region of real space and compute the localized orbitals of each *tessera* out of one specific pseudoeigenvalue equation to be solved in a basis set expansion approximation, using a basis set specific to the *tessera*, or orbital-specific basis set. In other words, a set of building-block embedded cluster pseudoeigenvalue coupled equations, one for each *tessera*, is solved in a self-consistent manner. Doing so, the method becomes parallel (the *tesserae* are computed independently of each other in the self-consistent procedure) and exhibits a linear-scaling dependence with the size of the molecule. We call the present method Mosaico.

Early works on localized orbitals proposed the ideas of computing them directly by a self-consistent procedure^{24,26} and computing one or several localized orbitals out of separate eigenvalue equations starting with a set of qualitatively localized orbitals.²⁷ These ideas have been

*e-mail luis.seijo@uam.es

†e-mail zoila.barandiaran@uam.es

used by several authors to propose methods leading to some particular sets of localized orbitals or to localized orbitals dependent on the initial guess.^{12,28,29,30,31} We apply them to the computation of the occupied localized orbitals of any localization method of choice. Here, we do not pay attention to the computation of virtual localized orbitals; methods for the determination of virtual localized orbitals useful in wave function based correlation methods have been used already in early works²⁸. The idea of using different basis sets for different regions of the system is also present in early work,²⁷ it has been used and discussed by several authors,^{5,12,31,32} and is a common procedure in embedded cluster and effective core potential calculations.^{20,21}

Besides its use as a linear scaling method in large molecules and solids, the Mosaico method can be used as an embedded cluster method, where only a few localized orbitals of a large system are treated variationally while the rest is taken from a calculation on a similar molecule and frozen, with obvious computational advantages. This can be done because the solutions of the building-block embedded cluster coupled equations are the localized orbitals corresponding to a given localization method of choice, which enables their transferability between similar molecules.

In Sec. II we present the details of the Mosaico method. We performed monitoring calculations on large poly(ethylen oxide) molecules and three dimensional carbon monoxide clusters using an extended Hückel semiempirical Hamiltonian, which are presented in Sec. III. They are aimed at showing the convergence of the parallel calculation to the right solution, the convergence of the total energy with the size of the orbital-specific basis sets towards the exact value, the linear-scaling characteristics of the method, and the performance of embedded cluster approximations.

II. METHOD

A. Basics of the method

Let us consider the Hartree-Fock equations in wave function based *ab initio* methods,^{18,33,34} the Kohn-Sham equations in density functional theory,³⁵ or the effective Hartree-Fock equations in semiempirical methods.³⁶ The canonical form of the spin-restricted closed-shell version of these equations can be written as

$$\hat{F}\underline{\varphi}^{can} = \underline{\varphi}^{can}\underline{\varepsilon}, \quad (1)$$

with an appropriate choice of the one-electron Hamiltonian \hat{F} for each case, where $\underline{\varphi}^{can}$ is a row vector of n occupied molecular orbitals,

$$\underline{\varphi}^{can} = (|\varphi_1^{can}\rangle, |\varphi_2^{can}\rangle, \dots, |\varphi_n^{can}\rangle), \quad (2)$$

and $\underline{\varepsilon}$ is an $n \times n$ diagonal matrix of orbital energies. This and what follows can be generalized to the spin-unrestricted cases if the orbitals and the Hamiltonian in

Eqs. 1 and 2 are adequately substituted by the α and β choices,³⁷ here we will continue with the detailed description of the spin-restricted closed-shell case for the sake of clarity. The virtual orbitals, $\underline{\varphi}^{vir}$, are also solutions of Eq. 1; in this paper, however, we will focus our attention on the occupied spectrum and, unless specified, we will refer only to different sets of occupied orbitals from now on.

The Fock-Dirac one-electron density operator is defined as

$$\hat{\rho} = \underline{\varphi}^{can}\underline{\varphi}^{can\dagger} = \sum_{i=1}^n |\varphi_i^{can}\rangle\langle\varphi_i^{can}|. \quad (3)$$

It is invariant under arbitrary unitary transformations of the occupied orbitals,

$$\underline{\varphi}^L = \underline{\varphi}^{can} \underline{canU}^L, \quad (4)$$

$$\hat{\rho} = \underline{\varphi}^{can}\underline{\varphi}^{can\dagger} = \underline{\varphi}^L\underline{\varphi}^{L\dagger}, \quad (5)$$

where we use the notation \underline{canU}^L for the unitary matrix that transforms canonical occupied orbitals onto localized orbitals of a given localization method, which we label L . Also, $\hat{\rho}$ is the projection operator of the occupied space,

$$\begin{aligned} \hat{\rho}\underline{\varphi}^{can} &= \underline{\varphi}^{can}, \\ \hat{\rho}\underline{\varphi}^L &= \underline{\varphi}^L, \\ \hat{\rho}\underline{\varphi}^{vir} &= 0. \end{aligned} \quad (6)$$

The fact that wave function, one-electron density, total energy, and Fock operator are invariant under unitary transformations within the occupied space (Eq. 4) has been exploited to define localized orbitals, which are useful to facilitate large scale calculations and to bridge the gap between extensive numerical calculations and qualitative chemical thinking. A localization method, say L , can be defined by its choice of \underline{canU}^L . Very sound and popular localization methods are the methods of Boys,²³ Edmiston-Ruedenberg,²⁴ and Pipek-Mezey,²⁵ although others have been proposed. All of the above can be used in first-principles methods; Pipek-Mezey's can also be applied in semiempirical methods. The common procedure to compute localized orbitals is to complete firstly a canonical calculation and, later, use the canonical orbitals in an iterative optimization process converging to \underline{canU}^L . Gilbert²⁶ has pointed out that any set of occupied localized orbitals formally fulfills the eigenvalue equation of an effective Fock (or Kohn-Sham) Hamiltonian defined as $\hat{F}^L = \hat{F} - \hat{\rho}\hat{F}\hat{\rho} + \hat{\rho}\hat{L}\hat{\rho}$,

$$\hat{F}^L\underline{\varphi}^L = [\hat{F} - \hat{\rho}\hat{F}\hat{\rho} + \hat{\rho}\hat{L}\hat{\rho}] \underline{\varphi}^L = \underline{\varphi}^L \underline{\lambda}, \quad (7)$$

where \hat{L} is a Hermitean localization operator and $\underline{\lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues of \hat{F}^L and of \hat{L} ,

$$\underline{\lambda} = \underline{\varphi}^{L\dagger}\hat{L}\underline{\varphi}^L. \quad (8)$$

Localization operators \hat{L} corresponding to the above mentioned localization methods are, in general, not known and Eq. 7 has not been exploited to compute the localized orbitals via diagonalization of the matrix $\underline{\varphi}^{can\dagger} \hat{L} \underline{\varphi}^{can}$, to our knowledge.

Eq. 7 has been discussed by several authors^{12,24,31} and it has been used as a basis for a building-block technique based on a self-consistent series of embedded cluster calculations.¹² The iterative solution of the building-block equations of Ref. 12, which contain arbitrary localization operators, leads to localized orbitals dependent on the initial guess and on the iteration procedure. Although in principle this is not detrimental for total energy calculations, it is an undesirable property because it creates problems of transferability of localized orbitals between similar molecules.

Here, on the basis of the same ideas than in Ref. 12, we present a new building-block and embedding method that, starting from an arbitrary guess and a choice of a particular localization method (Boys', Edmiston-Ruedenberg's, Pipek-Mezey's, or any other), leads, in a controlled manner, to the corresponding localized orbitals. We will call this method Mosaico.

Let us suppose we set the goal of computing the n occupied localized orbitals corresponding to a localization method L for the ground state of a molecule,

$$\underline{\varphi}^L = (|\varphi_1^L\rangle, |\varphi_2^L\rangle, \dots, |\varphi_n^L\rangle), \quad (9)$$

and related properties such as electron density and total energy. We see the whole set of localized orbitals as a mosaic of orbitals, and define subsystem, fragment, cluster, or *tesserae* as a subset of these orbitals which are localized in some region of real space. (For example, in an organic acid R-COOH we may define one of the subsystems or *tesserae* as that made of nine orbitals localized in the spatial region close to the COOH nuclei.) The terms subsystem, fragment, and cluster have been used by many authors with different meanings and we prefer to use the term *tessera* all over the paper for the present definition of subsystem. In this way, the whole mosaic of localized orbitals is made of N *tesserae* A, B, \dots , and the vector of localized orbitals can be rewritten as

$$\underline{\varphi}^L = (\underline{\varphi}_A^L, \underline{\varphi}_B^L, \dots), \quad (10)$$

where $\underline{\varphi}_A^L$ is a row vector with the n_A occupied orbitals of *tessera A*,

$$\underline{\varphi}_A^L = (|\varphi_{A1}^L\rangle, |\varphi_{A2}^L\rangle, \dots, |\varphi_{An_A}^L\rangle), \quad (11)$$

$\underline{\varphi}_B^L$ a row vector with the n_B orbitals of *tessera B*,

$$\underline{\varphi}_B^L = (|\varphi_{B1}^L\rangle, |\varphi_{B2}^L\rangle, \dots, |\varphi_{Bn_B}^L\rangle), \quad (12)$$

and so on, with $n_A + n_B + \dots = n$. The orbitals of a *tessera* define a subspace of the occupied space whose density and projection operator is

$$\hat{\rho}_A = \underline{\varphi}_A^L \underline{\varphi}_A^{L\dagger}. \quad (13)$$

The density operators of the *tesserae* fulfil

$$\hat{\rho}_A \hat{\rho}_B = \delta_{AB} \hat{\rho}_A, \quad (14)$$

$$\hat{\rho} = \sum_{B=1}^N \hat{\rho}_B, \quad (15)$$

$$\hat{\rho} \hat{\rho}_A = \hat{\rho}_A \hat{\rho} = \hat{\rho}_A, \quad (16)$$

where the sum in Eq. 15 extends over all the N *tesserae* of the system.

In the Mosaico method we seek to compute the localized orbitals of each *tessera* out of its own eigenvalue equation:

tessera A :

$$\hat{F}_A^L \underline{\varphi}_A^L = [\hat{F} - \hat{\rho} \hat{F} \hat{\rho} + \hat{\rho} \hat{L}_A \hat{\rho}] \underline{\varphi}_A^L = \underline{\varphi}_A^L \lambda_A, \quad (17)$$

tessera B :

$$\hat{F}_B^L \underline{\varphi}_B^L = [\hat{F} - \hat{\rho} \hat{F} \hat{\rho} + \hat{\rho} \hat{L}_B \hat{\rho}] \underline{\varphi}_B^L = \underline{\varphi}_B^L \lambda_B \quad (18)$$

...

($\lambda_A, \lambda_B, \dots$, being diagonal matrices of size $n_A \times n_A$, $n_B \times n_B, \dots$) under the conditions of Eqs. 13-16. These conditions are fulfilled if all the orbitals are eigenfunctions of the same Hermitean operator. In other words, all the orbitals $\underline{\varphi}_B^L$ for $B \neq A$ must be eigenfunctions of \hat{F}_A^L , all the orbitals $\underline{\varphi}_A^L$ for $A \neq B$ must be eigenfunctions of \hat{F}_B^L , and so on. This means that the subsystem localization operators $\hat{L}_A, \hat{L}_B, \dots$ of Eqs. 17, 18, ... must be such that the effective subsystem Hartree-Fock or Kohn-Sham Hamiltonians $\hat{F}_A^L, \hat{F}_B^L, \dots$, have the same eigenfunctions as \hat{F}^L in Eq. 7 but different eigenvalues. (Note that the virtual orbitals $\underline{\varphi}^{vir}$ are also eigenfunctions of Eqs. 17, 18, ... since $\hat{\rho} \underline{\varphi}^{vir} = 0$; their calculation has not been stated as a goal here and they will not be referred to in the rest of the Section.)

As commented above, the localization operator \hat{L} in Eq. 7 corresponding to a localization method L is usually not known. However, every localization method L has a well defined procedure for the computation of the unitary transformation matrix $_{can} \underline{U}^L$ (Eq. 4) in a given molecule.^{23,24,25} This procedure can also be applied to any orthogonal basis of the occupied space $\underline{\varphi}^{(0)}$ other than the canonical orbital basis $\underline{\varphi}^{can}$; the result is then the unitary matrix ${}_0 \underline{U}^L$ that transforms the initial non-canonical set of occupied orbitals onto the L -method localized orbitals,

$$\underline{\varphi}^L = \underline{\varphi}^{(0)} {}_0 \underline{U}^L. \quad (19)$$

At this point, we can express the operator \hat{L} of Eq. 7 as its spectral representation in any basis of the occupied space, e.g. $\underline{\varphi}^{(0)}$,

$$\hat{L} = \underline{\varphi}^L \lambda \underline{\varphi}^{L\dagger} = \underline{\varphi}^{(0)} {}_0 \underline{U}^L \lambda {}_0 \underline{U}^{L\dagger} \underline{\varphi}^{(0)\dagger}. \quad (20)$$

Now, consistently with the discussion following Eqs. 17, 18, ..., we can define the following Hermitean

subsystem or *tessera* localization operators:

$$\hat{L}_A = \underline{\varphi}^L \underline{\lambda}_{A(n)} \underline{\varphi}^{L\dagger} = \underline{\varphi}^{(0)} {}_0\underline{U}^L \underline{\lambda}_{A(n)} {}_0\underline{U}^{L\dagger} \underline{\varphi}^{(0)\dagger} \quad (21)$$

$$\hat{L}_B = \underline{\varphi}^L \underline{\lambda}_{B(n)} \underline{\varphi}^{L\dagger} = \underline{\varphi}^{(0)} {}_0\underline{U}^L \underline{\lambda}_{B(n)} {}_0\underline{U}^{L\dagger} \underline{\varphi}^{(0)\dagger} \quad (22)$$

...

where $\underline{\lambda}_{A(n)}, \underline{\lambda}_{B(n)}, \dots$ are diagonal matrices of size $n \times n$ with arbitrary diagonal real data. A choice consisting on low values for the n_A diagonal elements of $\underline{\lambda}_{A(n)}$ corresponding to the localized orbitals of *tessera* A , $\underline{\varphi}_A^L$, and sufficiently higher values for the $n - n_A$ remaining diagonal elements, guarantees that the localized orbitals of *tessera* A are computed as the lowest n_A eigenvectors of the effective Hartree-Fock or Kohn-Sham Hamiltonian \hat{F}_A^L (Eq. 17), which is a convenient choice for safe and efficient orbital selection in the iterations of self-consistent procedures. Obviously, the same comments stand for all *tesserae*. We may remark that a choice consisting on negative values for the n_A, n_B, \dots cited diagonal elements and zero values for the $n - n_A, n - n_B, \dots$ remaining elements, although not necessary, is efficient and simplifies the expression of the subsystem localization operators,

$$\hat{L}_A = \underline{\varphi}_A^L \underline{\lambda}_A \underline{\varphi}_A^{L\dagger} = \underline{\varphi}^{(0)} {}_0\underline{U}^{L(A)} \underline{\lambda}_A {}_0\underline{U}^{L(A)\dagger} \underline{\varphi}^{(0)\dagger} \quad (23)$$

$$\hat{L}_B = \underline{\varphi}_B^L \underline{\lambda}_B \underline{\varphi}_B^{L\dagger} = \underline{\varphi}^{(0)} {}_0\underline{U}^{L(B)} \underline{\lambda}_B {}_0\underline{U}^{L(B)\dagger} \underline{\varphi}^{(0)\dagger} \quad (24)$$

...

where $\underline{\lambda}_A$, of size $n_A \times n_A$, is the diagonal eigenvalue matrix of Eq. 17, and ${}_0\underline{U}^{L(A)}$ is a rectangular matrix made of n_A columns of ${}_0\underline{U}^L$, and similarly for all *tesserae*.

Using Eqs. 21, 22, ... in 17, 18, ..., plus the fact that $\hat{\rho}$ is the projection operator of the occupied space (Eq. 6), results in the working equations of the Mosaico method for a localization method of choice L :

tessera A :

$$\begin{aligned} \hat{F}_A^L \underline{\varphi}_A^L &= \left[\hat{F} - \hat{\rho} \hat{F} \hat{\rho} + \underline{\varphi}^{(0)} {}_0\underline{U}^L \underline{\lambda}_{A(n)} {}_0\underline{U}^{L\dagger} \underline{\varphi}^{(0)\dagger} \right] \underline{\varphi}_A^L \\ &= \underline{\varphi}_A^L \underline{\lambda}_A, \end{aligned} \quad (25)$$

tessera B :

$$\begin{aligned} \hat{F}_B^L \underline{\varphi}_B^L &= \left[\hat{F} - \hat{\rho} \hat{F} \hat{\rho} + \underline{\varphi}^{(0)} {}_0\underline{U}^L \underline{\lambda}_{B(n)} {}_0\underline{U}^{L\dagger} \underline{\varphi}^{(0)\dagger} \right] \underline{\varphi}_B^L \\ &= \underline{\varphi}_B^L \underline{\lambda}_B, \end{aligned} \quad (26)$$

...

Eq. 25 is the Mosaico equation for the embedded *tessera* A , Eq. 26 for the embedded *tessera* B , and so on. They are pseudoeigenvalue equations (even in the case of semiempirical methods with density-independent \hat{F} Hamiltonians.) They can be solved with standard SCF iterative procedures. Starting with an initial guess, $\underline{\varphi}^{(0)}$, the procedure of the localization method of choice, \underline{L} , is applied in order to compute ${}_0\underline{U}^L$, which, together with $\hat{\rho}$ and \hat{F} , give the embedded *tessera* Hamiltonians $\hat{F}_A^L, \hat{F}_B^L, \dots$ for the current iteration. (Note that the elements of the diagonal matrices $\underline{\lambda}_{A(n)}, \underline{\lambda}_{B(n)}, \dots$, are input real numbers.) Solving the eigenvalue equations leads to new

orbitals which are used to update $\underline{\varphi}^{(0)}$ and iterate. At convergence, $\underline{\varphi}^{(0)} = \underline{\varphi}^L$ and ${}_0\underline{U}^L$ is the unit matrix. Also, the eigenvalues of the *tesserae* are identical to the input values of $\underline{\lambda}_{A(n)}, \underline{\lambda}_{B(n)}, \dots$. For instance, the n_A non-zero values of $\underline{\lambda}_A$ coincide with the non-zero values of $\underline{\lambda}_{A(n)}$ associated with the localized orbitals of *tessera* A .

Several options are open for iterative procedures leading to the solutions of the Mosaico equations 25, 26, ... They all have to face two basic types of iterations: (1) The microiterations, or *tessera* iterations, which are standard SCF iterations addressed to solve one of the embedded *tessera* equations, e.g. Eq. 25, for fixed orbitals of the other *tesserae*. All the methods available for speeding the solution of the canonical Hartree-Fock or Kohn-Sham equations can be used here. (2) The macroiterations, or mosaic iterations, which involve repeated solutions of all the embedded *tessera* equations. The macroiterations can be performed sequentially on a given list of *tesserae*, e.g. A, B, \dots, A, B, \dots , meaning that the orbitals computed for *tessera* A are used in the calculation of *tessera* B , and so on. Most interestingly, they can be performed in parallel, meaning that the calculations of all *tesserae* A, B, \dots are done using the whole mosaic of orbitals from the previous macroiteration. This alternative is very important, because it allows to take full advantage of parallelism at a high level of the calculation. In other words, the full Mosaico calculation is made of a sequence of macroiterations, each of them consisting of a parallel set of Hartree-Fock, Kohn-Sham, or semiempirical calculations on the embedded *tesserae* A, B, \dots . Each of these *tessera* calculations, which can be time consuming, can be performed in a separate processor or computer.

Alternately, the macroiterations can be done *à la carte*, that is restricted to only one or several selected *tesserae*. This is to say that only some of the localized orbitals are optimized whereas the rest of them are frozen. This is the embedded cluster approximation. Its reliability rests upon the transferability of the localized orbitals. Embedded cluster calculations are significantly cheaper than the full calculation of a complete system. They can be performed on a molecule or solid if a calculation on a similar molecule or solid has been carried out before in order to provide the orbitals of the embedding frozen *tesserae*. They are specially useful to study defects in solids, chemisorption, series of molecules with different substituents, or reactions taking place in local regions of large molecules.

The Mosaico equations (Eqs. 25, 26, ...), which give the localized orbitals of a given localization method L and, in consequence, the same total energy and electron density than the canonical calculation, are the basis for approximations that make them useful in practice. These approximations, which resort to truncations supported by the localized nature of the orbitals, are systematic and converge to the exact result. They are basically twofold: (1) An orbital-specific basis set approximation can be adopted, where the localized orbitals of a *tessera* are expanded in a different basis set than the localized orbitals

of another *tessera*. This can be done because the n_A occupied localized orbitals of *tessera A* are the only orbitals to be computed by solving equation 25, whereas the $n - n_A$ orbitals of the other *tesserae* are discarded; consequently, a local basis set can be used as long as it is sufficient for a good representation of the localized orbitals $\underline{\varphi}_A^L$. Obviously, this is true for any *tessera*. (2) A localization algorithm based on local rotations can be used in each iteration instead of the standard algorithm of the localization method of choice, L . In the local rotations algorithm, the localized orbitals of *tessera A* are computed in each iteration using a subset of the $\underline{\varphi}^{(0)}$ orbitals which is localized in *tesserae* not too distant from *A*. This can be done because the target localized orbitals are computed out of input localized orbitals rather than out of canonical orbitals.

B. Use of orbital-specific basis sets, OSBS

In a standard molecular calculation, all the orbitals of a molecule are expanded in a common basis set, which is the basis set of the molecule and consists of n_{BSF} basis set functions,

$$\underline{\chi} = (|\chi_1\rangle, |\chi_2\rangle, \dots, |\chi_{n_{BSF}}\rangle), \quad (27)$$

which could be contracted Gaussian functions or other sort of local functions. In the orbital-specific basis set approximation, OSBS, the localized orbitals of *tessera A* are expanded in a local basis set made of b_A functions, the localized orbitals of *tessera B* in a local basis set made of b_B functions, and so on. Although not necessary, it is very convenient that the local basis sets are subsets of the global basis set, and that several *tesserae* share a number of basis set functions. The local basis sets can be represented by the following row vectors,

tessera A :

$$\underline{\chi}_A = (|\chi_{A1}\rangle, |\chi_{A2}\rangle, \dots, |\chi_{Ab_A}\rangle), \quad (28)$$

tessera B :

$$\underline{\chi}_B = (|\chi_{B1}\rangle, |\chi_{B2}\rangle, \dots, |\chi_{Bb_B}\rangle), \quad (29)$$

...

Accordingly, eqs. 25, 26, ..., take the usual matrix form,¹⁸

$$\text{iessera A} : \underline{F}_A^L \underline{C}_A^L = \underline{S}_A \underline{C}_A^L \underline{\lambda}_A \quad (30)$$

$$\text{iessera B} : \underline{F}_B^L \underline{C}_B^L = \underline{S}_B \underline{C}_B^L \underline{\lambda}_B, \quad (31)$$

...

In Eq. 30, for instance, \underline{F}_A^L and \underline{S}_A are the $b_A \times b_A$ matrix representations of the \hat{F}_A^L and the unit operators in the $\underline{\chi}_A$ basis set,

$$\underline{F}_A^L = \underline{\chi}_A^\dagger \hat{F}_A^L \underline{\chi}_A, \quad \underline{S}_A^L = \underline{\chi}_A^\dagger \underline{\chi}_A, \quad (32)$$

and \underline{C}_A^L is the $b_A \times n_A$ matrix of localized orbital coefficients,

$$\underline{\varphi}_A^L = \underline{\chi}_A \underline{C}_A^L. \quad (33)$$

The solution of Eq. 30 can be attained using standard $\mathcal{O}(b_A^3)$ diagonalization procedures. Except for low gap materials, where the degree of localization attainable is limited and the localized orbitals may decay slowly with distance,³⁸ the local basis set size b_A is expected to remain within reasonable limits for a diagonalization or, more generally, for an embedded cluster calculation. Obviously, the growth of b_A will impose more practical limitations in 3D systems, like bulk solids and very big clusters, than in 2D and 1D systems, like many molecules. In general, all the methods useful to speed up a standard molecular calculation, like convergence acceleration methods, can be used with Eq. 30. But, in this case, additional advantage can be taken from the fact that only a small number of eigenvalues/eigenfunctions are computed for each *tessera*, and efficient diagonalization algorithms used in CI calculations, like the multiroot Davidson-Liu method,^{39,40} can be applied to significantly reduce the prefactor of the $\mathcal{O}(b_A^3)$ dependence.

Let us comment on the calculation of the effective Hamiltonian matrix of a *tessera*, \underline{F}_A^L , where the localized nature of the orbitals is also profitable. For simplicity, we will call $\underline{\phi}$ the row vector of the n current localized orbitals ($\underline{\varphi}^L$, Eq. 19) at a given iteration, which can be written as the union of the row vectors $\underline{\phi}_A, \underline{\phi}_B, \dots$, that contain the n_A, n_B, \dots current localized orbitals of *tessera A, B, ...* respectively,

$$\underline{\phi} = (\underline{\phi}_A, \underline{\phi}_B, \dots). \quad (34)$$

The $b_A \times b_A$ effective Hamiltonian matrix of *tessera A* is

$$\begin{aligned} \underline{F}_A^L &= \underline{\chi}_A^\dagger \hat{F}_A^L \underline{\chi}_A \\ &= \underline{\chi}_A^\dagger \hat{F} \underline{\chi}_A - \underline{\chi}_A^\dagger \hat{\rho} \hat{F} \hat{\rho} \underline{\chi}_A + \underline{\chi}_A^\dagger \underline{\phi} \underline{\lambda}_{A(n)} \underline{\phi}^\dagger \underline{\chi}_A \end{aligned} \quad (35)$$

The first term in the right hand side of Eq. 35 is a diagonal block of the Hamiltonian matrix of the full system. The second term can be expanded in inter-*tessera* terms by inserting Eq. 15:

$$\begin{aligned} \underline{\chi}_A^\dagger \hat{\rho} \hat{F} \hat{\rho} \underline{\chi}_A &= \\ \sum_{B=1}^N \sum_{C=1}^N & \left(\underline{\chi}_A^\dagger \underline{\phi}_B \right) \left(\underline{\phi}_B^\dagger \hat{F} \underline{\phi}_C \right) \left(\underline{\phi}_C^\dagger \underline{\chi}_A \right). \end{aligned} \quad (36)$$

In the evaluation of Eq. 36, we can take advantage of the locality of basis sets and molecular orbitals by evaluating a Δ_{AB}^S interaction table and a Δ_{AB}^F interaction table. Δ_{AB}^S is set to 0 if all the elements of the $b_A \times n_B$ matrix $\underline{\chi}_A^\dagger \underline{\phi}_B$ and all the elements of the $b_B \times n_A$ matrix $\underline{\phi}_B^\dagger \underline{\chi}_A$ have an absolute value lower than a given threshold, and is set to 1 otherwise. Similarly, Δ_{AB}^F is set to 0 if all the elements of the $n_A \times n_B$ matrix $\underline{\phi}_A^\dagger \hat{F} \underline{\phi}_B$ have an absolute

value lower than a given threshold, and is set to 1 otherwise. Rearranging Eq. 36 and using these interaction tables, we can write

$$\begin{aligned} \underline{\chi}_A^\dagger \hat{\rho} \hat{F} \hat{\rho} \underline{\chi}_A &= \sum_{B=1}^N \Delta_{AB}^S (\underline{\chi}_A^\dagger \underline{\phi}_B) (\underline{\phi}_B^\dagger \hat{F} \underline{\phi}_B) (\underline{\phi}_B^\dagger \underline{\chi}_A) \\ &+ \sum_{B=2}^N \Delta_{AB}^S \sum_{C=1}^{B-1} \Delta_{AC}^S \Delta_{BC}^F [(\underline{\chi}_A^\dagger \underline{\phi}_B) (\underline{\phi}_B^\dagger \hat{F} \underline{\phi}_C) (\underline{\phi}_C^\dagger \underline{\chi}_A) \\ &+ \text{adjoint}] . \end{aligned} \quad (37)$$

The use of the *tesserae* interaction tables Δ_{AB}^S and Δ_{AB}^F guarantees that the calculation of the present term does not scale as N^2 because the number of *tesserae* in the neighborhood of (or interacting with) *tessera A* does not increase indefinitely with the size of the molecule. Also, note that the interaction tables do not need to be updated every macroiteration, because the localized orbitals do not experience big changes in size after a few iterations. It should be noticed that we assume that a linear-scaling method is used for the computation of the Hamiltonian matrix of the full system,^{4,7,41} of which diagonal and non-diagonal blocks are needed in Eqs. 35 and 37.

The last term in Eq. 35 can be written as

$$\begin{aligned} \underline{\chi}_A^\dagger \underline{\phi} \underline{\lambda}_{A(n)} \underline{\phi}^\dagger \underline{\chi}_A &= \\ \sum_{B=1}^N \Delta_{AB}^S (\underline{\chi}_A^\dagger \underline{\phi}_B) \underline{\lambda}_{A(n)B} (\underline{\phi}_B^\dagger \underline{\chi}_A) , \end{aligned} \quad (38)$$

where the Δ_{AB}^S interaction table is used and $\underline{\lambda}_{A(n)B}$ is the $n_B \times n_B$ diagonal matrix resulting from the elements of $\underline{\lambda}_{A(n)}$ corresponding to the localized orbitals of *tessera B*. For the particular choice of the arbitrary diagonal matrix $\underline{\lambda}_{A(n)}$ leading to Eq. 23, Eq. 38 is further simplified:

$$\begin{aligned} \underline{\chi}_A^\dagger \underline{\phi} \underline{\lambda}_{A(n)} \underline{\phi}^\dagger \underline{\chi}_A &= \underline{\chi}_A^\dagger \underline{\phi}_A \underline{\lambda}_A \underline{\phi}_A^\dagger \underline{\chi}_A \\ &= \underline{\chi}_A^\dagger \left(\sum_{i=1}^{n_A} |\phi_A\rangle \lambda_{Ai} \langle\phi_A| \right) \underline{\chi}_A \end{aligned} \quad (39)$$

where all λ_{Ai} must be negative.

The computation of the \underline{F}_A^L matrices (Eqs. 35, 37, and 38) and their diagonalization (Eq. 30) can be performed in parallel. This is graphically indicated in the upper part of Fig. 1. The global scaling of this diagonalization step is $\sum_{B=1}^N \mathcal{O}(b_B^3)$; in the case of all *tesserae* having the same local basis set size, \bar{b} , and same number of interactions with other *tesserae*, the scaling is $\mathcal{O}(\bar{b}^3 N)$.

C. Localization by local rotations

Usually, a given set of localized orbitals of a molecule, $\underline{\varphi}^L$, is computed out of the canonical orbitals (Eq. 4). Well defined algorithms are routinely used to calculate the $n \times n$ unitary transformation matrix $\underline{\text{can}}U^L$, which

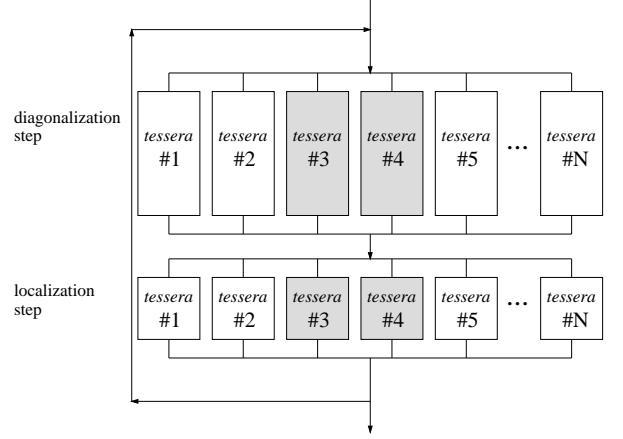


FIG. 1: Diagram of the two basic parallel loops of a macroiteration. In the first loop, the embedded *tesserae* effective Hamiltonian matrices are computed and diagonalized (Eqs. 30, 31, ...). In the second loop, the ${}_0\underline{U}_{CA}^L$ matrices of the localization method of choice, L , are computed out of the orbitals resulting from the first loop and the localization transformations (Eqs. 30, 31, ...) are performed. The highlighted boxes would be the only to be entered in a embedded cluster calculation where *tesserae* #3 and #4 define the active cluster.

are normally of order n^3 or higher.^{24,25} In the present method, however, they are computed in each macroiteration out of other set of localized orbitals (Eq. 19). We can expect the contributions of the initial localized orbitals $\underline{\varphi}^{(0)}$ to the target localized orbitals $\underline{\varphi}^L$ to decay with distance. Accordingly, a Δ_{AB}^{LR} local rotation table can be computed, where $\Delta_{AB}^{LR} = 1$ if the initial localized orbitals of *tessera B* are used to compute the target localized orbitals of *tessera A* (and viceversa) and $\Delta_{AB}^{LR} = 0$ otherwise. Several options to compute the Δ_{AB}^{LR} local rotation table are possible. Reasonable choices are to use the same criterium as for the Δ_{AB}^F interaction table except for the use of a different threshold or, simply, make the Δ_{AB}^{LR} local rotation table identical to the Δ_{AB}^F interaction table.

Arranging the initial localized orbitals in *tesserae*,

$$\underline{\varphi}^{(0)} = (\underline{\varphi}_A^{(0)}, \underline{\varphi}_B^{(0)}, \dots) , \quad (40)$$

the approximation of local rotations for the localization step can be written as

$$\text{tessera } A : \quad \underline{\varphi}_A^L = \sum_{C=1}^N \Delta_{AC}^{LR} \underline{\varphi}_C^{(0)} {}_0\underline{U}_{CA}^L , \quad (41)$$

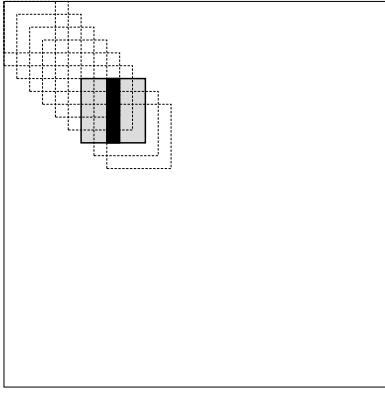


FIG. 2: Schematic representation of the localization by local rotations (step 2 in Fig. 1). The unitary matrix ${}_0\mathbf{\underline{U}}^L$ is represented together with the blocks that are computed in the localization parallel loop for *tesserae* #1 to #10. The block computed for *tessera* #8 (${}_0\mathbf{\underline{U}}_{LR}^L$ in Eq. 43) is highlighted. The subblock representing the actual localized orbitals of this *tessera* is represented with a rectangular fill box.

$$\text{tessera } B : \quad \underline{\varphi}_B^L = \sum_{C=1}^N \Delta_{BC}^{LR} \underline{\varphi}_C^{(0)} {}_0\mathbf{\underline{U}}_{CB}^L, \quad (42)$$

...

In Eq. 41, ${}_0\mathbf{\underline{U}}_{CA}^L$ is the $n_C \times n_A$ block of $\mathbf{\underline{U}}^L$ with the contributions of the initial localized orbitals of *tessera* C to the target localized orbitals of *tessera* A . The column block made of all ${}_0\mathbf{\underline{U}}_{CA}^L$ with $\Delta_{AC}^{LR} = 1$ is represented by the fill rectangular box shown in Fig. 2. This column block can be calculated approximately as the column block of the submatrix of ${}_0\mathbf{\underline{U}}^L$ indicated in Fig. 2 by the open square box. We can call this submatrix ${}_0\mathbf{\underline{U}}_{LRA}^L$, so that if $\underline{\varphi}_{LRA}^{(0)}$ is a row vector with the initial localized orbitals of all *tesserae* C having $\Delta_{AC}^{LR} = 1$ (of length $n_{LRA} = \sum_{C=1}^N \Delta_{AC}^{LR} n_C$), then

$$\underline{\varphi}_{LRA}^L = \underline{\varphi}_{LRA}^{(0)} {}_0\mathbf{\underline{U}}_{LRA}^L. \quad (43)$$

${}_0\mathbf{\underline{U}}_{LRA}^L$ can be computed by simple application of the localization method of choice to the set of n_{LRA} initial localized orbitals $\underline{\varphi}_{LRA}^{(0)}$. Note that only the n_A orbitals of $\underline{\varphi}_{LRA}^L$ corresponding to *tessera* A ($\underline{\varphi}_A^L$) are taken from Eq. 43. Lowering the threshold of the local rotation table improves the precision of this approximation systematically.

The computation of the ${}_0\mathbf{\underline{U}}_{LRA}^L$ matrices (Eqs. 43) can be performed in parallel. This is graphically indicated in the lower part of Fig. 1. The global scaling of this localization step is $\sum_{B=1}^N \mathcal{O}(n_{LRB}^\ell)$, where ℓ is the order

of the localization method of choice L (e.g. 3 in Pipek-Mezey and 5 in Edmiston-Ruedenberg methods); if the same number of orbitals are used in the local rotations of all *tesserae*, \bar{n}_{LR} , the scaling is $\mathcal{O}(\bar{n}_{LR}^\ell N)$.

Since the localization by local rotations among initial localized orbitals goes together with the use of orbital-specific basis sets, the resulting localized orbitals of *tessera* A must be represented with the basis set $\underline{\chi}_A$ alone. This can be achieved by truncation of $\underline{\varphi}_A^L$ after the transformation (Eq. 41) or, alternatively, by its projection on the $\underline{\chi}_A$ space, $\underline{\chi}_A (\underline{\chi}_A^\dagger \underline{\chi}_A)^{-1} \underline{\chi}_A^\dagger \underline{\varphi}_A^L$. We have not observed any practical advantage in projecting instead of truncating, neither in the precision attainable in the total energy nor in the convergence, in all the tests performed.

D. Symmetry

The orbitals resulting from the present method do not belong to irreducible representations of the molecular point symmetry group because they are not eigenfunctions of the totally symmetric one-electron Hamiltonian \hat{F} . However, except for symmetry breaking localization methods, they are related by the symmetry operations of the molecule and this fact can be used to reduce computing time.⁴² In effect, all *tesserae* orbitals can be obtained by applying molecular symmetry operations, \hat{R} , to a list of symmetry independent *tesserae*, so that if A is a symmetry independent *tessera* and *tessera* B is obtained from A by symmetry operation \hat{R} , then $\underline{\varphi}_B^L = \hat{R} \underline{\varphi}_A^L$. In other words, the diagonalizations and localizations described in Sections II B and II C can be performed only on the list of symmetry independent *tesserae*. The computation of F_A^L (Eq. 35) can also profit from this property.

Also, the site symmetry of the *tesserae* can be used in exactly the same way that molecular symmetry is used in standard molecular calculations, because the matrix of the effective Hamiltonian of any embedded *tessera* is blocked according to the irreducible representations of the local point symmetry group of that particular *tessera*.

Besides the use of symmetry, one can also take advantage of quasisymmetry for speeding up purposes. So, if two *tesserae* A and B are quasisymmetry related by an operation \hat{R}_q ($\underline{\varphi}_B^L \approx \hat{R}_q \underline{\varphi}_A^L$), they can be treated as symmetry related for a number of macroiterations, in which the list of *tesserae* where the diagonalization and localization steps are performed can be shortened, finally releasing all quasisymmetry restrictions up to a full convergence of the Mosaico calculation.

E. Summary of the Mosaico algorithm

In summary, the algorithm for a Mosaico calculation of a molecule is the following:

1. Take the current guess of localized orbitals of the

molecule and do, in parallel, for each symmetry independent *tessera*:

- (a) compute the embedded *tessera* effective Hamiltonian matrix (Eqs. 35, 37, and 38),
- (b) diagonalize it (Eq. 30) and compute new *tessera* orbitals (Eq. 33),

2. Take all the orbitals produced in step 1 and do, in parallel, for each symmetry independent *tessera*:

 - (a) compute the ${}_0\mathcal{U}_{LR_A}^L$ unitary matrix (Eq. 43) corresponding to the localization method of choice L by applying the corresponding localization algorithm, and take the columns that correspond to the current *tessera*,
 - (b) compute the target *tessera* localized orbitals (Eq. 41),

 3. Check for convergence and iterate on step 1 if necessary. Compute properties upon convergence.

The parallel steps 1 and 2 are schematically represented in Fig. 1.

III. RESULTS

In this section, we present the results of monitoring calculations on poly(ethylene oxide) molecules, $\text{H}(\text{CH}_2\text{OCH}_2)_m\text{H}$, (Sec. III A) and three dimensional CO clusters, $(\text{CO})_m$, (Sec. III B) aimed at showing the convergence of the parallel calculations to the right solutions, the convergence of the total energies with the size of the orbital-specific basis sets towards the exact values, and the linear-scaling of the method. We also include embedded cluster calculations on defective systems resulting from a chemical substitution of one O atom by a S atom in poly(ethylene oxide), (Sec. III C) aimed at showing the performance of embedded cluster calculations vs. full system calculations.

All the calculations are single point energy calculations with an Extended Hückel (EH) Hamiltonian.³⁶ This is a convenient choice to monitor the Mosaico procedure because, on the one hand, the Hamiltonian and overlap matrices of the EH method are isomorphous with their *ab initio* counterparts, and on the other, the Hamiltonian is not self-consistent and its computation is straightforward. In this way, the analysis of timing and scaling is focused on the orbital optimization or diagonalization part and free from contaminations due to the computation of the Hamiltonian matrix. The calculations have been performed with a 2GB RAM personal computer with the program Mosaico;⁴³ the Extended Hückel Hamiltonian and overlap matrices have been calculated with the program EHT.⁴⁴ Although the loops performed mimic parallel loops (Fig. 1), the total elapsed times shown hereafter correspond to sequential loops performed in a single processor. In the present version of the program we have

paid special attention to the scaling features, whereas the prefactors are highly improvable. This fact, together with the average performance of the personal computer used, makes the absolute values of the elapsed times shown in this section of little value; instead, it is the scaling of the method what is relevant.

A. Poly(ethylene oxide)

Poly(ethylene oxide) (PEO)⁴⁵ is a polymer widely used in the field of polymer electrolytes of molecular formula $\text{H}(\text{CH}_2\text{OCH}_2)_m\text{H}$. The molecular calculations we present for different numbers of monomers, m , use the localization method of projected localized molecular orbitals (PLMO)⁴⁶ (see Appendix) using a very simple set of reference orbitals: $s_A + s_B$ for all A-B pairs of bonded atoms, plus $p_y(O) + p_z(O)$ and $p_y(O) - p_z(O)$ for all oxygen atoms, where y and z are local cartesian axis on the oxygen assuming the C-O-C atoms define a xy plane with the y axis bisecting the C-O-C angle.

In all these calculation we used the following definition for the *tesserae* or subsystems: One *tessera* made of 10 orbitals localized in the sigma bonds and lone pairs of $\text{CH}_3\text{OCH}_2^-$ [which, in the localization method used for these particular calculations, are the 10 orbitals with one-to-one maximum overlap with the reference orbitals $s(H_1) + s(C_1)$, $s(C_1) + s(H_2)$, $s(C_1) + s(H_3)$, $s(C_1) + s(O_1)$, $p_y(O_1) + p_z(O_1)$, $p_y(O_1) - p_z(O_1)$, $s(O_1) + s(C_2)$, $s(C_2) + s(H_4)$, $s(C_2) + s(H_5)$, $s(C_2) + s(C_3)$], plus $m - 2$ *tesserae* made of 9 orbitals localized in the sigma bonds and lone pairs of the next $\text{CH}_2\text{OCH}_2^-$ groups, and a final *tessera* of 9 orbitals localized in the sigma bonds and lone pairs of the terminal CH_2OCH_3 .

We performed the Mosaico calculations using three different orbital-specific basis sets: In the calculations labeled 1N, the orbitals of each *tessera* have been represented with a subset of the global basis set consisting of all the basis set functions of the atoms involved in the bonds and lone pairs of the *tessera* plus those of the atoms involved in bonds and lone pairs of the first neighbor *tesserae* or monomers. In the 2N and 3N calculations, the orbital-specific basis sets were extended to second and third neighbor monomers. All the calculations converge to the same results regardless of the initial guess and the iteration procedure (parallel or any kind of sequential choice). The 9 localized orbitals which constitute one of the bulk *tesserae* of $\text{H}(\text{CH}_2\text{OCH}_2)_{30}\text{H}$ are shown in Fig. 3.

Fig. 4 shows the total energy of $\text{H}(\text{CH}_2\text{OCH}_2)_{30}\text{H}$ as a function of the orbital-specific basis sets used, which converges to the exact energy in the limit of a standard calculation where all orbitals are spanned in a common basis set. As the size of the OSBS increases, the lack of full orthogonality between orbitals of different *tesserae* originated by the basis set truncation becomes negligible and, accordingly, the difference between the total energy properly computed and the total energy computed un-

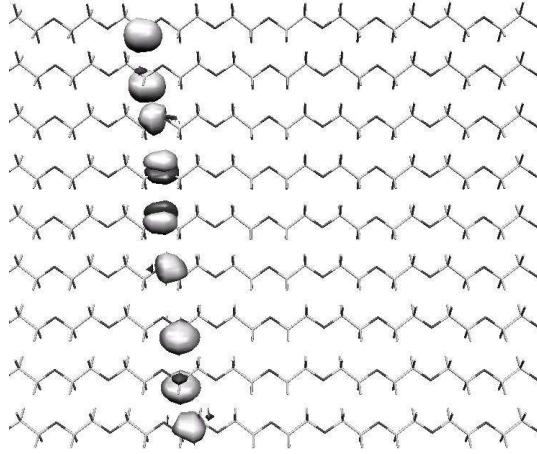


FIG. 3: The nine localized orbitals that constitute one of the bulk *tesserae* used in $\text{H}(\text{CH}_2\text{OCH}_2)_{30}\text{H}$.

der the assumption that the orbitals are fully orthogonal vanishes. The energy loss per monomer due to the use of orbital-specific basis sets instead of a common basis set for all orbitals, is presented in Table I for several orbital-specific basis sets as a function of the molecular size. The predictability of the energy losses is apparent. Fig. 5 shows the convergence with macroiterations of the total energy of the $\text{H}(\text{CH}_2\text{OCH}_2)_{30}\text{H}$ polymer at the 2N level of orbital-specific basis set.

Fig. 6 shows the wall clock elapsed time per macroiteration in the calculation of the $\text{H}(\text{CH}_2\text{OCH}_2)_m\text{H}$ molecules as a function of the number of monomers m (which in this case coincides with the number of *tesserae*, N), of atoms, and of basis set functions. The times per macroiteration and *tessera* spent in the diagonalization

TABLE I: Energy loss per monomer with respect to canonical calculations, in hartree/monomer, of poly(ethylen oxide) $\text{H}(\text{CH}_2\text{OCH}_2)_m\text{H}$ molecules and $(\text{CO})_m$ clusters, as calculated with several orbital-specific basis sets.

	OSBS			$(E - E_{\text{canonical}})/m$
$\text{H}(\text{CH}_2\text{OCH}_2)_m\text{H}$	$m = 10$	$m = 20$	$m = 50$	
1N	1.14×10^{-4}	1.28×10^{-4}	1.37×10^{-4}	
2N	1.74×10^{-7}	2.09×10^{-7}	2.30×10^{-7}	
3N	$< 1 \times 10^{-11}$	$< 1 \times 10^{-11}$	2.0×10^{-10}	
$(\text{CO})_m$	$m = 13$	$m = 63$		
1N	3.56×10^{-7}	7.74×10^{-7}		
2N	2.34×10^{-7}	5.13×10^{-7}		
3N	3.3×10^{-8}	3.7×10^{-8}		

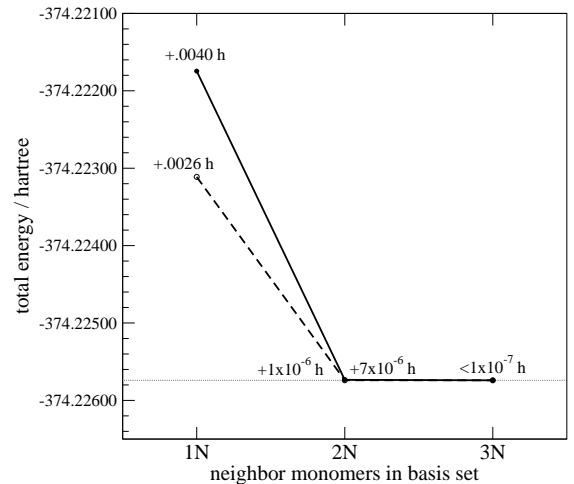


FIG. 4: Total energy of $\text{H}(\text{CH}_2\text{OCH}_2)_{30}\text{H}$ as a function of the orbital-specific basis set size. 1N, 2N, and 3N labels indicate that each *tessera* is calculated with the basis set functions of all atoms up to first, second, and third neighbor monomers, respectively. Energy losses with respect to the exact canonical energy are indicated on the lines, in hartree units. Full line: correct calculation of the total energy. Dashed line: total energy calculated under the assumption of perfect orthogonality between the localized orbitals.

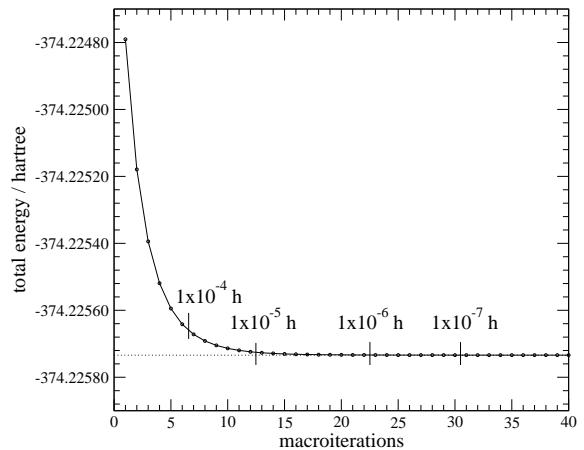


FIG. 5: Convergence with macroiterations of the total energy of $\text{H}(\text{CH}_2\text{OCH}_2)_{30}\text{H}$ with 2N orbital-specific basis set. Convergence to several subunits of hartree are indicated. A similar number of macroiterations for convergence has been found in all poly(ethylen oxide) polymers studied.

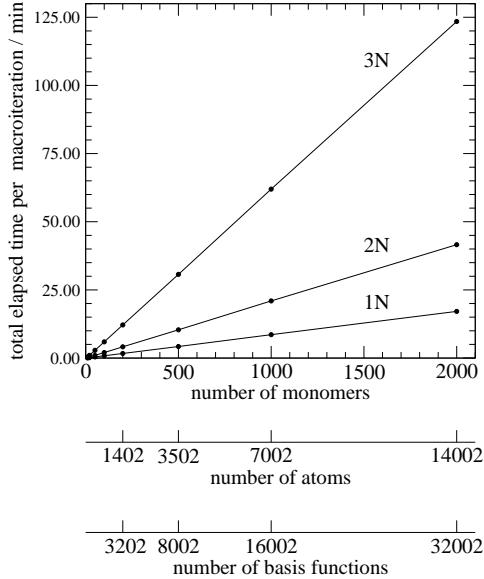


FIG. 6: Total wall clock elapsed times per macroiteration in the calculation of poly(ethylen oxide) molecule $\text{H}(\text{CH}_2\text{OCH}_2)_m\text{H}$ ($m = 10, 20, 50, 100, 200, 500, 1000, 2000$).

step and in the localization step are shown in Fig. 7. The horizontal lines reflect the $\mathcal{O}(N)$ scalings of both steps. The $\mathcal{O}(\langle b^3 \rangle)$ scaling of the diagonalization step is clearly reflected in the separation between the 1N, 2N, and 3N horizontal lines. The drop of the lines at low number of monomers is due to surface effects: The ratio between edge *tesserae* and bulk *tesserae* is significant in small polymers and, since edge *tesserae* are less demanding in terms of orbital-specific basis set and in terms of number of inter-*tesserae* interactions, they reduce the computing time with respect to what it would be if all of them were bulk *tesserae*. As we will see below, this fortunate surface effect, which lowers time with respect to a set of N bulk *tesserae*, is much more pronounced in 3D systems. The localization times shown in the bottom part of Fig. 7 are very similar in the 1N, 2N and 3N calculations. This is so because they depend basically on the number of orbitals used in the local rotations, which is the same in the three calculations. For the particular localization method we used for these calculation, PLMO, the elapsed times scale as $\sum_{B=1}^N \mathcal{O}(n_{LRB}^3) \approx \mathcal{O}(\langle n_{LRB}^3 \rangle) \mathcal{O}(N)$. The small dependence with the size of the orbital-specific basis sets is related with the lengths of the matrix transformations in Eq. 47.

B. $(\text{CO})_m$ clusters

We performed Mosaico calculations on three dimensional $(\text{CO})_m$ clusters of several sizes, extracted from

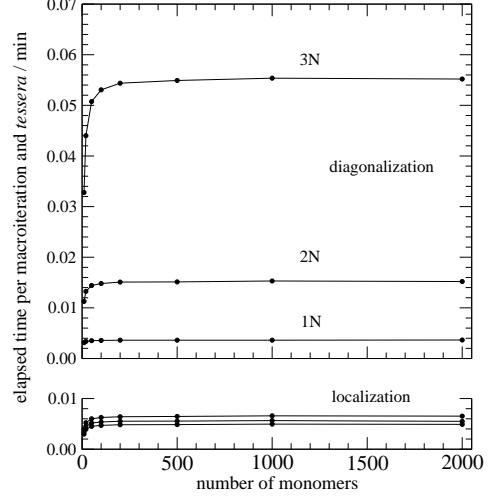


FIG. 7: Elapsed times per macroiteration and *tessera* in the calculation of poly(ethylen oxide) molecule $\text{H}(\text{CH}_2\text{OCH}_2)_m\text{H}$. Up: Diagonalization step. Down: Localization step.

the crystalline structure of the α phase of solid carbon monoxide ($P2_13$ spatial group, CO bond length $r(\text{C}-\text{O}) = 1.128 \text{ \AA}$, cell constant $a_0 = 5.64 \text{ \AA}$),⁴⁷ which is an interesting material known to experience irreversible photopolymerization under pressure.^{48,49} In this crystal, a bulk CO molecule has a first-neighbor coordination number 12 (CO molecules at a distance between centers of gravity of 3.99 \AA), and second- and third-neighbor coordination numbers 18 and 42 (CO molecules at 5.64 \AA and 6.91 \AA , respectively). The $(\text{CO})_{63}$ cluster is represented in Fig. 8 as an example.

In these calculations we also used the PLMO localization method, with the reference orbitals defined as the $5m$ valence occupied canonical orbitals of the m isolated CO molecules. We defined each *tessera* to be made of 5 orbitals localized in the spatial region of a CO molecule, which, for the chosen localization method, means the 5 localized orbitals with maximum overlap with the canonical orbitals of the CO molecule. The orbital-specific basis sets used have been labeled 1N, 2N, and 3N when the basis set of a *tessera* consists of the basis set functions of its C and O atoms plus the basis set functions of the atoms of the first, second, and third-neighbor CO molecules, respectively (see above).

The energy losses per CO molecule (Table I) are small and diminish as the size of the OSBS increases. Times per macroiteration and *tessera* spent in the diagonalization step and in the localization step are shown in Fig. 9. The diagonalization times spent in a inner or

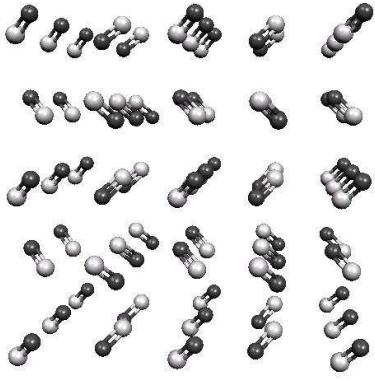


FIG. 8: $(CO)_{63}$ cluster as a piece of the α phase of solid carbon monoxide (space group $P2_13$).

bulk *tesserae*, which is the most demanding, are included as a reference. This time is constant for all clusters where the inner *tessera* is a truly bulk *tessera*, that is, where its OSBS and its interaction tables die off before the limits of the cluster; this situation is reached in a smaller cluster for the 1N OSBS and in a bigger cluster for the 3N OSBS. Naturally, the times spent per *tessera* in the full system calculations are lower than the times spent in the bulk *tesserae*, the differences showing the importance of surface effects: smaller clusters and bigger orbital-specific basis sets have a larger ratio of surface/bulk *tesserae* and, accordingly, a larger time reduction with respect to the bulk *tesserae*. In a large m regime, the bulk *tesserae* are dominant and set the asymptotic limit of the full system times, which scale linearly with the cluster size. The $\mathcal{O}(\langle b^3 \rangle)$ dependence is shown by the asymptotic values of the 1N, 2N, and 3N lines, as well as by the values of $\langle b^3 \rangle = (\sum_{A=1}^m b_A^3)/m$ printed along the 3N line in Fig. 9. The localization times (bottom of Fig. 9) are very much independent of the size of the OSBS because they are directly dependent in the number of occupied orbitals included in the local rotations, which is the same in 1N, 2N, and 3N calculations; the small dependence shown in the Figure is due to the fact that the transformations down to the basis set level depend on the OSBS size.

C. Embedded cluster calculations

The Mosaico method can be used for embedded cluster calculations, where the computational effort is focused on an active site of a molecule, comprising only a number of relevant *tesserae*, while the rest of it is taken from a previous calculation on a similar molecule and frozen. In this section we show the results of embedded cluster calculations on $H(CH_2OCH_2)_p-CH_2SCH_2-(CH_2OCH_2)_pH$. $H(CH_2OCH_2)_p-CH_2SCH_2-(CH_2OCH_2)_pH$ can be re-

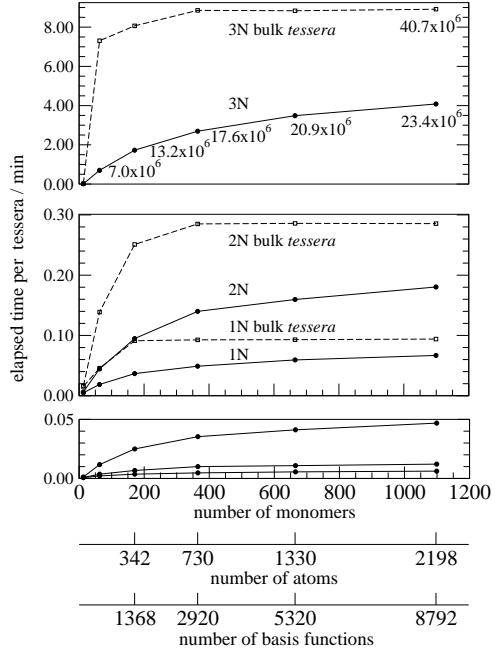


FIG. 9: Elapsed times per macroiteration and *tessera* spent in the calculation of $(CO)_m$ clusters. Down: Localization step of the 1N, 2N, and 3N OSBS calculations. Middle: Diagonalization step of the 1N and 2N OSBS calculations. Up: Diagonalization step of the 3N OSBS calculations; The values of $\langle b^3 \rangle = (\sum_{A=1}^m b_A^3)/m$, where b_A is the number of basis set functions in the OSBS of *tessera* A, are indicated.

garded as the result of creating a chemical defect in $H(CH_2OCH_2)_mH$, with $p = (m - 1)/2$, by substitution of the central oxygen by a sulfur atom. We may expect the localized orbitals distant from the S atom in the “defective” molecule to be very similar to the orbitals of the “perfect” molecule localized in the same spatial regions and, accordingly, we may take them from a previous calculation on $H(CH_2OCH_2)_mH$ and use them in a Mosaico calculation of $H(CH_2OCH_2)_p-CH_2SCH_2-(CH_2OCH_2)_pH$ where they are kept frozen; this defines an embedded cluster Mosaico calculation. Canonical and Mosaico calculations on the $m = 21$ polymer (using the same nuclear configurations before and after the creation of the S defect) reveal that the precision reached with a 3N orbital-specific basis set on the perfect polymer requires a better OSBS after creating this chemical defect: A 5N basis set for the central defective *tessera* and its 5 neighbor *tesserae* together with a 3N basis set for the remaining *tesserae*, gives an energy error of 4.5×10^{-9} hartree/monomer. Taking this into account, we performed embedded cluster calculations using a 5N OSBS for the variational *tesserae* and taking the orbitals that remain frozen for the rest of *tesserae* in the polymer from the 3N OSBS calculation on the “perfect” polymer

molecule $\text{H}(\text{CH}_2\text{OCH}_2)_m\text{H}$. The total energy errors of these embedded cluster calculations (with respect to the full molecular Mosaico calculations) on the polymers of 21 and 201 monomers are shown in Fig. 10 as a function of the active cluster size (number of active *tesserae*).

It is shown that the errors are too large to be acceptable if only the central defective *tessera* is active, whereas they drop to an acceptable value of 40×10^{-6} hartree when the orbitals of the first-neighbor *tesserae* are optimized as well, and to less than 1×10^{-6} hartree when second-neighbor *tesserae* are part of the variational embedded cluster. These errors are the same in the two $m = 21$ and $m = 201$ polymers, as corresponds to the local nature of the chemical defect. The times spent in the embedded cluster calculations, as a fraction of the times spent in the respective full system calculations, are shown at the bottom of Fig. 10. Overall, this figure illustrates the potentiality of embedded cluster calculations where the transferability of the localized orbitals of a localization method of choice is exploited.

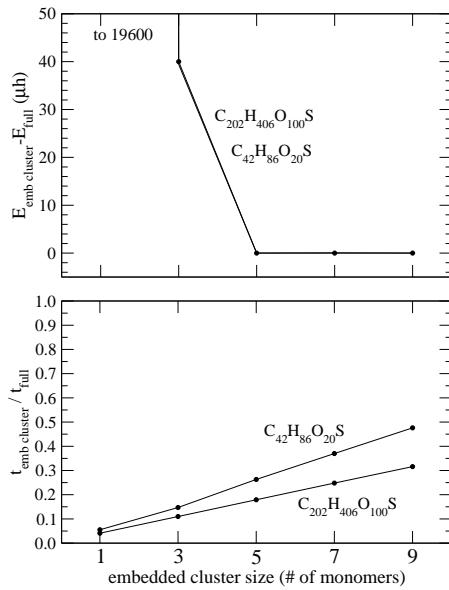


FIG. 10: Embedded cluster calculations on $\text{H}(\text{CH}_2\text{OCH}_2)_p\text{-CH}_2\text{SCH}_2\text{-(CH}_2\text{OCH}_2)_p\text{H}$ ($p = 10, 50$) with a 5N OSBS for the variational *tessera* and 3N OSBS for the remaining *tesserae*. The abscissa labels correspond to the number of variational *tesserae* included in the active clusters; the defective *tessera* is always the central one. Above: Total energy errors, in $\mu\text{hartree}$, with respect to full molecule Mosaico calculations using 5N OSBS in the 11 central *tesserae* and 3N OSBS in the rest; the result of the smallest cluster is indicated. Below: Elapsed time per macroiteration, as a fraction of the time per macroiteration of the full molecule Mosaico calculation.

IV. CONCLUSIONS

We presented a new linear-scaling method for the energy minimization step of semiempirical and first-principles Hartree-Fock and Kohn-Sham calculations, which we abbreviated under the name Mosaico. In this method, a set of embedded *tessera* pseudoeigenvalue coupled equations is solved in a building-block self-consistent fashion, which results in optimum occupied localized orbitals of any localization method of choice, represented with orbital-specific basis sets. The Mosaico method is parallel at a high level of the calculation. It can be used in full system calculations as well as in embedded cluster calculations, where only an active fraction of the localized molecular orbitals of the whole system are variational while the rest are taken from a similar molecule and kept frozen.

We presented the results of monitoring, single point energy calculations with the extended Hückel Hamiltonian on poly(ethylene oxide) molecules and three dimensional carbon monoxide clusters with very large number of basis set functions. Total energy losses due to the use of orbital-specific basis sets are small for reasonably small sizes of these and total energies converge to the canonical values when the orbital-specific basis sets are increased towards the limit of a common basis set for all the localized orbitals. Convergence of total energy with self-consistent macroiterations is good and elapsed times per macroiterations have been shown to scale linearly with the molecular size. Besides the full system calculations, the good performance of the much less demanding embedded cluster approach has been shown in total energy calculations on defective systems resulting from chemical substitution of an oxygen atom by a sulfur atom in poly(ethylene oxide) molecules. The transferability of the localized orbitals of a given localization method between similar molecules has been shown to lead to the same total energy precision than full molecular calculations at a fraction of the computational cost.

Acknowledgments

This research was supported in part by Dirección General de Investigación, Ministerio de Ciencia y Tecnología, Spain, under contract BQU2002-01316, and Ministerio de Educación, Cultura y Deportes, Spain, Accciones de Movilidad PR2003-0024 and PR2003-0027. We are very grateful to Professor Martin Head-Gordon, University of California, Berkeley, and to his research group for their hospitality.

APPENDIX

A. Projected localized molecular orbitals

A simple localization method has been proposed by Ruedenberg *et al.*,⁴⁶ which has not reached the popularity of other methods like Boys,²³ Edmiston-Ruedenberg,²⁴ and Pipek-Mezey²⁵ methods. Although it has been formulated in a context of atoms-in-molecules, it is a general method. Here, we reformulate it in the general case and in the case of localization by local rotations (Sec. II C), using the present notation.

Given a set of n input occupied (canonical or localized) orthogonal orbitals of a molecule, $\underline{\varphi}^{(0)} = (|\varphi_1^{(0)}\rangle, |\varphi_2^{(0)}\rangle, \dots, |\varphi_n^{(0)}\rangle)$, and a set of n reference arbitrary orbitals, $\underline{\xi} = (|\xi_1\rangle, |\xi_2\rangle, \dots, |\xi_n\rangle)$, the n projected localized orthogonal orbitals, $\underline{\varphi}^{PLMO} = (|\varphi_1^{PLMO}\rangle, |\varphi_2^{PLMO}\rangle, \dots, |\varphi_n^{PLMO}\rangle)$, are defined as those resulting from a unitary transformation of $\underline{\varphi}^{(0)}$ which are most similar to $\underline{\xi}$. The PLMOs correspond to maximizing the one-to-one overlaps with the reference orbitals $\underline{\xi}$, or the functional $\sum_{i=1}^n |\langle \varphi_i^{PLMO} | \xi_i \rangle|$, under orthogonality constraints⁵⁰ and can be computed as^{46,50}

$$\underline{\varphi}^{PLMO} = \hat{\rho} \underline{\xi} \left[\underline{\xi}^\dagger \hat{\rho} \underline{\xi} \right]^{-\frac{1}{2}} = \underline{\varphi}^{(0)} {}_0 \underline{U}^{PLMO}, \quad (44)$$

with

$${}_0 \underline{U}^{PLMO} = (\underline{\varphi}^{(0)\dagger} \underline{\xi}) \left[(\underline{\varphi}^{(0)\dagger} \underline{\xi})^\dagger (\underline{\varphi}^{(0)\dagger} \underline{\xi}) \right]^{-\frac{1}{2}}. \quad (45)$$

In the approximation of local rotations (Section II C), the projected localized orbitals of a *tessera A* are computed out of n_{LRA} ($n_{LRA} > n_A$) input localized orbitals $\underline{\varphi}_{LRA}^{(0)}$ and n_{LRA} reference orbitals $\underline{\xi}_{LRA}$, both of them including the localized/reference orbitals belonging to the *tesserae* included in the local rotations, with the expression:

$$\underline{\varphi}_A^{PLMO} = \underline{\varphi}_{LRA}^{(0)} {}_0 \underline{U}_{LRA}^{PLMO}, \quad (46)$$

being

$${}_0 \underline{U}_{LRA}^{PLMO} = \left\{ \left(\underline{\varphi}_{LRA}^{(0)\dagger} \underline{\xi}_{LRA} \right) \times \left[(\underline{\varphi}_{LRA}^{(0)\dagger} \underline{\xi}_{LRA})^\dagger (\underline{\varphi}_{LRA}^{(0)\dagger} \underline{\xi}_{LRA}) \right]^{-\frac{1}{2}} \right\}_{n_A-\text{col}}, \quad (47)$$

where it has been indicated that only the n_A columns that correspond to *tessera A* are computed and used.

Among the advantages of the PLMO localization method are its speed and its simplicity, because the usual iterative optimization procedures involved in localization^{23,24,25} are substituted by a one-step calculation of the reciprocal square root of a symmetrical matrix, which is an $\mathcal{O}(n^3)$ process (or $\mathcal{O}(n_{LRA}^3)$ in local rotations). Its main disadvantage is the requirement of an external, arbitrary set of reference orbitals, $\underline{\xi}$. Although this is a limitation in calculations of reactivity, it is not a practical problem in molecular structure calculations where the nature of the bonds is known in advance and finding good reference bond and lone pair orbitals is not difficult. We may remark that the application of the PLMO method with a reference consisting of a given set of localized orbitals, e.g. Edmiston-Ruedenberg's, leads exactly to that set of orbitals. This property can be exploited in many ways and, in particular, in order to remove the arbitrariness inherent to the PLMO method. For instance, it can be used to produce reference orbitals for the PLMO method (to be used in large molecules) out of Edmiston-Ruedenberg's or other non-arbitrary localized sets computed in selected sets of relatively small molecules. Also, a Mosaico calculation addressed to produce a given set of localized orbitals like Edmiston-Ruedenberg's can be safely performed using the chosen localization method in some macroiterations and the faster PLMO method, with the current ER orbitals as a reference set, in the rest of them.

- ¹ S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).
- ² P. Ordejón, Phys. Status Solidi B **217**, 335 (2000).
- ³ C. A. White and M. Head-Gordon, J. Chem. Phys. **101**, 6593 (1994).
- ⁴ C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, Chem. Phys. Lett. **253**, 268 (1996).
- ⁵ P. Ordejón, E. Artacho, and J. M. Soler, Phys. Rev. B **53**, 10441 (1996).
- ⁶ M. C. Strain, G. E. Scuseria, and M. J. Frisch, Science **271**, 5245 (1996).
- ⁷ G. E. Scuseria, J. Phys. Chem. A **103**, 4782 (1999).
- ⁸ C. Ochsenfeld, C. A. White, and M. Head-Gordon, J. Chem. Phys. **109**, 1663 (1998).
- ⁹ W. Yang, Phys. Rev. Lett. **66**, 1438 (1991).

- ¹⁰ W. Yang and T. S. Lee, J. Chem. Phys. **103**, 5674 (1995).
- ¹¹ T. S. Lee, D. M. York, and W. Yang, J. Chem. Phys. **105**, 2744 (1996).
- ¹² L. Seijo and Z. Barandiarán, J. Math. Chem. **10**, 41 (1992).
- ¹³ P. Ordejón, D. A. Drabold, R. M. Martin, and M. P. Grumbac, Phys. Rev. B **51**, 1456 (1995).
- ¹⁴ J. P. Stewart, Int. J. Quantum Chem. **58**, 133 (1996).
- ¹⁵ A. Shukla, M. Dolg, and H. Stoll, Phys. Rev. B **58**, 4325 (1998).
- ¹⁶ T. Helgaker, H. Larsen, J. Olsen, and P. Jorgensen, Chem. Phys. Lett. **327**, 397 (2000).
- ¹⁷ M. Head-Gordon, Y. Shao, C. Saravanan, and C. A. White, Mol. Phys. **101**, 37 (2003).
- ¹⁸ C. C. J. Roothaan, Rev. Mod. Phys. **23**, 69 (1951).

¹⁹ Z. Barandiarán and L. Seijo, *J. Chem. Phys.* **89**, 5739 (1988).

²⁰ L. Seijo and Z. Barandiarán, in *Computational Chemistry: Reviews of Current Trends*, edited by J. Leszczyński (World Scientific, Singapore, 1999), vol. 4, p. 55.

²¹ J. L. Whitten and H. Yang, *Surf. Sci. Rep.* **24**, 55 (1996).

²² N. Govind, Y. A. Wang, and E. A. Carter, *J. Chem. Phys.* **110**, 7677 (1999).

²³ S. F. Boys, *Rev. Mod. Phys.* **32**, 296 (1960).

²⁴ C. Edmiston and K. Ruedenberg, *Rev. Mod. Phys.* **35**, 457 (1963).

²⁵ J. Pipek and P. G. Mezey, *J. Chem. Phys.* **90**, 4916 (1989).

²⁶ T. L. Gilbert, in *Molecular Orbitals in Chemistry, Physics, and Biology*, edited by P. O. Löwdin and B. Pullman (Academic, New York, 1964), pp. 405–420.

²⁷ D. Peters, *J. Chem. Phys.* **51**, 1559 (1969).

²⁸ D. L. Wilhite and J. L. Whitten, *J. Chem. Phys.* **58**, 948 (1973).

²⁹ A. B. Kunz, *J. Phys. B* **6**, L47 (1973).

³⁰ H. Schlosser, *Chem. Phys. Lett.* **23**, 545 (1973).

³¹ O. Matsuoka, *J. Chem. Phys.* **66**, 1245 (1977).

³² H. Stoll, G. Wagenblast, and H. Preuss, *Theor. Chim. Acta* **57**, 169 (1980).

³³ V. Fock, *Z. Physik* **61**, 126 (1930).

³⁴ D. R. Hartree and W. Hartree, *Proc. Roy. Soc. A* **150**, 9 (1935).

³⁵ W. Kohn and L. J. Sham, *Phys. Rev.* **140**, 1133 (1965).

³⁶ R. Hoffmann, *J. Chem. Phys.* **39**, 1397 (1963).

³⁷ J. A. Pople and R. K. Nesbet, *J. Chem. Phys.* **22**, 571 (1954).

³⁸ W. Kohn, *Phys. Rev. B* **7**, 4388 (1973).

³⁹ E. R. Davidson, *J. Comput. Phys.* **17**, 87 (1975).

⁴⁰ B. Liu, Technical Report LBL-8158, Lawrence Berkeley Laboratory, University of California, Berkeley, 1978.

⁴¹ J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, *J. Phys. Cond. Mat* **14**, 2745 (2002).

⁴² D. Peters, *Theor. Chim. Acta* **24**, 16 (1972).

⁴³ Mosaico 1.0 is a program for building-block and embedding calculations based on the use of localized orbitals and orbital-specific basis sets, written by L. Seijo, Departamento de Química, Universidad Autónoma de Madrid, Madrid, Spain.

⁴⁴ EHT 1.0 is a program for Extended Hückel Theory written by T. Liu and D. G. Truhlar, Department of Chemistry, University of Minnesota, U.S.A.

⁴⁵ F. M. Gray, *Solid Polymer Electrolytes* (VCH, New York, 1991).

⁴⁶ K. Ruedenberg, M. W. Schmidt, and M. M. Gilbert, *Chem. Phys.* **71**, 51 (1982).

⁴⁷ B. O. Hall and H. M. James, *Phys. Rev. B* **13**, 3590 (1976).

⁴⁸ A. I. Katz, D. Schiferl, and R. L. Mills, *J. Phys. Chem.* **88**, 3176 (1984).

⁴⁹ S. Bernard, G. L. Chiarotti, S. Scandolo, and E. Tosatti, *Phys. Rev. Lett.* **81**, 2092 (1998).

⁵⁰ E. Francisco, L. Seijo, and L. Pueyo, *J. Solid State Chem.* **63**, 391 (1986).